AD 697 403

STATISTICAL INFORMATION RETRIEVAL SYSTEM

Nicholas M. DiFondi

Rome Air Development Center
Griffiss Air Force Base, New York

October 1969

STATISTICAL INFORMATION RETRIEVAL SYSTEM

Nicholas M. DiFondi

FOREWORD

Research described in this report was accomplished under Project 4594, Task 459401.

This document has been reviewed by the Information Office (EMLS) and is releasable to the Clearinghouse for Scientific and Technical Information.

This report has been reviewed and is approved.

Francis M Kelly

Approved: JOHN A. THOMPSON, Lt Col, USAF
Chief, Intel Data Handling Branch
Intelligence & Reconnaissance Division

A E Stoll

Approved: A. E. STOLL, Colonel, USAF
Chief, Intel & Recon Division

FOR THE COMMANDER: Irving J Gabelman
IRVING J. GABELMAN
Chief, Plans Office

ABSTRACT

An information retrieval system was developed using technical word occurrences as a basis for classification.

A set of words, designated a vocabulary, was selected from the middle range of a frequency listing of words occurring in an experimental sample of 94 documents. The selection produced 115 non-function words with technical definition that did not allow ambiguous usage and they were assigned one of eighty concept numbers. The frequencies of these concepts served as data for factor analysis and 39 factors were extracted to represent the orthogonal axes of a geometric subject-content space. The locations of concepts in this space were used to locate the geometric position of documents according to their frequencies in the documents.

The total of 194 documents was used in the measuring of system effectiveness. The Mahalanobis D^2 function provided a statistical measure of the separation between relevant and not-relevant groups in the space. Linear discriminant functions were solved to maximize between group differences and Fisher's variance ratio was used to test the significance of group separation. Empirical and theoretical probabilities of misclassification were compared and system error on the average was .2% for relevant and 4.1% for not-relevant documents. Theoretical errors were 96.9% and 3.1% respectively. The small system errors

validated the accuracy of the 39 dimensional subject-content space.

Requests formulated for a previous experiment using the same data base were processed. Precision and Recall measures were calculated and on the average 66% Precision and 80% Recall were attained with one of three dissemination thresholds.

Overall analysis of the results supports the theory that statistical data about word occurrences is sufficient to accurately represent documents relative to their subject content.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. PURPOSE

The theory tested is whether a classification scheme developed from information concerning the occurrences of technical words can accurately represent documents relative to their subject content.

2. APPROACH

The number of occurrences of pre-selected technical words in a sample of the data base was factor analyzed. The result was a multi-dimensional classification space where the location of documents was an indication of subject content. Access to the space for retrieval purposes was gained by locating requests in the same manner as documents. The proximity of documents to the request was the basis for determining relevance.

3. EXPERIMENTAL DESIGN

3.1 DATA BASE COLLECTION

The data base was organized to reflect the interests of personnel employed in the Information Processing Branch of RADC who served as test subjects for this experiment. Articles representing 23 subject fields were selected from information processing journals. From these articles, parts were extracted according to specific content as follows:

1. An abstract or section which best represents the article's content
2. A section about the methods involved
3. A section about results
4. A section not representative of the main theme or article content

These sections were used to represent the journal articles and are referred to as documents in the data base. Although all sections were not present in some articles, the 194 documents collected contained as a minimum an abstract (or a suitable substitute section) and as a maximum all of the above sections.

3.2 VOCABULARY SELECTION

Vocabulary selection plays a crucial role in the development of this type of retrieval system where the final content representation of documents and search requests is dependent on the vocabulary terms occurring in each.

For this experiment, a random sample of 94 documents was selected from the data base to develop a classification scheme. The following procedures were used to select terms from the experimental sample. First a frequency program listed all single words and their total occurrences from most to least frequent. Then to obtain maximum discriminant power between subject fields, terms were selected from the middle range of frequencies. High frequencies are associated with words used in most of the fields whereas low frequencies are indicative of words highly specific to a minute portion of a total field. The middle range boundaries were arbitrarily selected since they are functions of individual data bases and the total number of different words appearing in them. Frequencies of 115 and 19 were chosen. Words within these boundaries were selected for the vocabulary if they had technical definition which was not so broad that usage could be ambiguous.

Complying with all boundaries and restrictions, the frequency listing yielded 80 single terms. Further analysis of words that occurred less than 20 times revealed certain forms (plurals, etc.) of the words chosen for the vocabulary which should be included also. For example, the selected word "handwritten" in a document about handwritten character recognition may also be found in such forms as "written" and "handwriting" due to style or the tense of the sentence in which it would appear. Consequently, those forms of vocabulary words considered desirable were included if the sum of their frequency and the original word's frequency did not exceed the upper middle range boundary of 115. This increased the total number of vocabulary words to 115. The 80 single terms were assigned concept numbers from 1 to 80 and the attached forms were given the same concept number as the original.

3.3 FREQUENCY DATA

A program called WORD MATCH then processes the 194 documents and counts the occurrences of each vocabulary word in each document. These word frequencies form a 194 x 80 document-term matrix and those frequencies associated with the experimental sample are then processed by program FACTOR ANALYSIS.

3.4 FACTOR ANALYSIS

At this point the content of each document is represented by the number of occurrences of each concept in the document. The minimum residual method of factor analysis extracted 39 factors which were rotated to a varimax solution of simple structure (Harmon; Lawley and Maxwell). Consequently, each word is represented by a 39 element vector where each value is the relationship the word has to that factor.

3.5 DOCUMENT LOCATIONS

The 39 factor loadings for any vocabulary word are best described as the geometric representations of the position of that word in a 39 dimensional orthogonal space. The 39 geometric representations necessary to describe the position of a document in this space can be derived from the loadings of words which occur in that document. One method of calculating document locations (Ossorio) averages respective factors across words and weights this by taking into account the consistency of these factor loadings. Only that part which calculates the average of the word factor loadings was used so that frequency data was the only influence on system effectiveness.

$$F_{iD_j} = \frac{1}{N}\sum_{k=1}^{N} F_{iw_k}$$

where

F_{iD_j} = the $i^{\underline{th}}$ factor loading for the $j^{\underline{th}}$ document.

F_{iw_k} = the $i^{\underline{th}}$ factor loading of the $k^{\underline{th}}$ vocabulary word occurring in the $j^{\underline{th}}$ document.

N = the number of vocabulary words occurring in the $j^{\underline{th}}$ document.

4. SYSTEM EFFECTIVENESS MEASURE

There are various methods of measuring the effectiveness of a retrieval system (Swets). Some require the formulation of requests to effect retrieval, and require knowledge of the total number of documents relevant to each request in order to compute effectiveness.

4.1 REQUEST Problem

SIRS as such does not allow for a straight-forward method of request formulation. In some systems such as SMART (Salton), the request is simply a few sentences which state the user's needs. Others require the user or an information specialist to transform the request into a set of keywords which best represent the user's needs. The user or specialist may also be allowed to weight these keywords according to relative importance At first glance, this type of request formulation would seem suitable to SIRS where the weights are frequency estimates. However, at this point in time such weights would be strictly speculative. The method of retrieval through keywords is tested and results are presented in Section 9.

4.2 RELEVANT DOCUMENT SELECTION

A method of evaluating the effectiveness of SIRS without benefit of retrieval was developed through multivariate statistical analysis techniques. In a previous experiment for which this data base had been compiled, 9 engineers were asked to formulate requests (sentence type) pertinent to their areas of interest resulting in 19 inquiries. The engineers were asked to manually search the data base and select those documents which satisfy each of their requests.

Assuming the remaining documents to be not-relevant, the effectiveness of the system should be a function of the distance between the two sets of documents in the space.

4.3 MAHALANOBIS D^2

The geometric representation of a document consists of 39 factor loadings which form a vector from the origin to the point which locates the document in the space. The user has decided which documents are relevant to his request so that it can be represented by a corresponding number of relevant document vectors. The remaining documents in the data base represent a set of not-relevant document vectors. A mean document vector can be calculated for each set simply by averaging corresponding factors. A mean difference vector results when corresponding factor loadings of the mean relevant document vector and the mean not-relevant document vector are subtracted. The following notation was used:

$\mathbf{X}_i^R$ = set of relevant document vectors.

$\mathbf{X}_i^{\bar{R}}$ = set of not-relevant document vectors.

$\bar{\mathbf{X}}^R$ = mean relevant document vector.

$\bar{\mathbf{X}}^{\bar{R}}$ = mean not-relevant document vector.

N_R = size of the relevant set of documents

$N_{\bar{R}}$ = size of the not-relevant set of documents.

Then

$$\mathbf{d} = \bar{\mathbf{x}}^R - \bar{\mathbf{x}}^{\bar{R}}$$

is the mean difference vector.

There is a variance associated with each set of documents given by

$$\mathbf{S}_R = \frac{1}{N_R}\left[\sum_{i=1}^{N_R} \mathbf{x}_i^R \mathbf{x}_i^{R'} - N_R \bar{\mathbf{x}}^R \bar{\mathbf{x}}^{R'}\right]$$

(prime notation indicates matrix transpose).

$$\mathbf{S}_{\bar{R}} = \frac{1}{N_{\bar{R}}}\left[\sum_{i=1}^{N_{\bar{R}}} \mathbf{x}_i^{\bar{R}} \mathbf{x}_i^{\bar{R}'} - N_{\bar{R}} \bar{\mathbf{x}}^{\bar{R}} \bar{\mathbf{x}}^{\bar{R}'}\right]$$

S_R and $S_{\bar{R}}$ are variance-covariance matrices for each set of documents where the diagonal elements are pure variances and the off diagonal elements are covariances. The two matrices are combined tc form a pooled variance-covariance matrix using the following formula.

$$\mathbf{S} = \frac{(N_R - 1)\mathbf{S}_R + (N_{\bar{R}} - 1)\mathbf{S}_{\bar{R}}}{N_R + N_{\bar{R}} - 2}$$

Mahalanobis D^2 is a multivariate statistical method for calculating the square of the distance between two samples (Li).

$$D^2 = \mathbf{d}'\mathbf{S}^{-1}\mathbf{d}$$

4.4 STATISTICAL SIGNIFICANCE

The Mahalanobis D^2 value can be tested to determine if the distance between the means of the two groups is statistically significant. Fisher's variance ratio F can be used to test significance (Li).

$$F = \frac{N-p-1}{(N-2)p} T^2 \qquad N = N_R + N_{\bar{R}}$$

where T^2 is Hotelling's generalization of Student's t^2 and is

$$T^2 = \frac{N_R \; N_{\bar{R}}}{N_R + N_{\bar{R}}} D^2$$

All the notation except for p has been defined previously.

p = the number of variables considered (39 factors).

The degrees of freedom associated with F are

$$\nu_1 = p = 39$$

$$\nu_2 = N - p - 1 = 194 - 39 - 1 = 154$$

ν_1 is the degrees of freedom for the numerator of F and ν_2 is the degrees of freedom for the denominator.

Of 19 requests, D^2 was found to be significant in 14 cases at the .05 level with 12 of these significant at the .01 level. These results are interpreted to mean that in 14 requests, the mean of the relevant set of documents is significantly different from the mean of the not-relevant set of documents based on their respective factor loadings with error of 5% or less. As a minimum, the system is based on information which is useful in most cases in distinguishing between two groups of documents. However, more meaningful measures are needed to extend the concept of significance to retrieval effectiveness.

5. LINEAR DISCRIMINANT FUNCTION.

In the previous section the variance ratio F is calculated as a direct function of D^2 by substitution of the expression for the generalized T^2.

$$F = \frac{N - p - 1}{(N-2)p} \cdot \frac{N_R N_{\bar{R}}}{N} \cdot D^2$$

The first ratio is constant across requests because neither N nor p change. The second ratio limits the value of F and requires large distances between groups when large disparities in size exist. It is possible for a small group of relevant documents to be imbedded in a large group of not-relevant documents and still have a statistically significant distance between their means. Thus, the groups overlap or intermingle and it becomes difficult to distinguish members of the smaller group from some members of the larger group. A transformation must be performed on the data such that maximum difference is achieved between the two groups.

Discriminant function procedures are such that the measures on the variables are combined to produce maximum differences between groups.

$$L = a_1 X + a_2 Y$$

is a linear discriminant function for a two dimensional (or variable) case. The coefficients are calculated in such a manner that maximum difference is exhibited through the discriminant scores (L values). In the present situation the linear discriminant function would be of the form

$$L_1 = a_1 X_1 + a_2 X_2 + \; . \quad . \quad . + a_{39} X_{39}$$

where the X's are the 39 factor loadings of the first document. The linear discriminant function describes a line with 194 points on it representing the discriminant scores calculated for the data base. The following formulas can be used to calculate the mean discriminant score for each group in order to circumvent the calculation of all discriminant scores.

$$L_R = (\mathbf{S}^{-1}\mathbf{d})' \bar{\mathbf{x}}^R \qquad\qquad L_{\bar{R}} = (\mathbf{S}^{-1}\mathbf{d})' \bar{\mathbf{x}}^{\bar{R}}$$

The discriminant function increases discriminating power between groups without changing the distance between their means since

$$D^2 = L_R - L_{\bar{R}}$$

5.1 PROBABILITIES OF MISCLASSIFICATION.

Under the assumption of normality and equal variances, the two groups may be geometrically represented along the linear discriminant function as shown in figure 1.

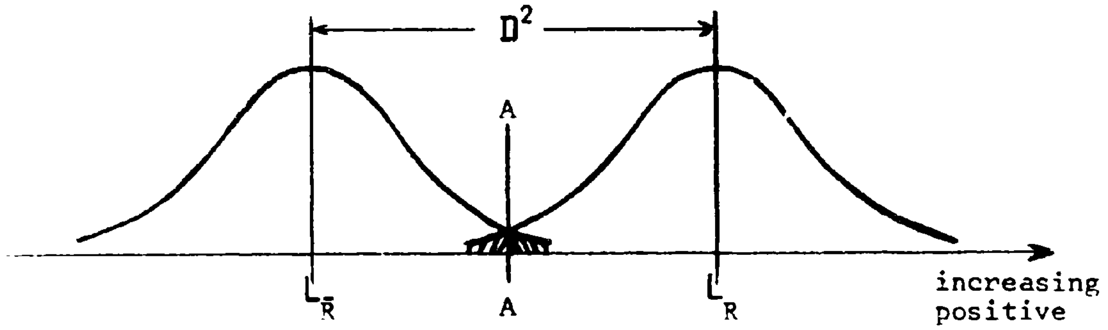

Fig. 1

Relevant and Not-Relevant Groups Projected on the Linear Discriminant Function

The line AA is the midpoint between the two group means. If any a priori probability is not introduced, line AA would represent the decision boundary for classification into one of the two groups. The probability of misclassification would be equal for both and is represented by the shaded portion in Figure 1.

The conventional procedure used to reduce the probability of misclassification in linear discriminant analysis when classifying into one of two groups is to introduce a loss or cost function, or both. (The cost of misclassification is usually difficult to assess and will not be considered). The relationship used in reduction is dependent upon the mean discriminant value associated with each group. For example, in this system $L_R > L_{\bar{R}}$ (see fig. 1), then the loss function is the log of the ratio of the a priori probabilities of the groups

$$\ln(P_R \;/\; P_{\bar{R}})$$

where

$$P_{\bar{R}} = N_{\bar{R}}/N \qquad\qquad P_R = N_R/N$$

The natural logarithm is used because each group has a probability density function which is exponential to the base e. The location of line AA (fig. 1) on the discriminant function is defined as

$$(L_R + L_{\bar{R}})/2$$

and misclassification is reduced by subtracting the loss function

$$\left[(L_R + L_{\bar{R}})/2\right] - \ln(P_R \;/\; P_{\bar{R}})$$

This expression relocates the decision boundary such that misclassification is decreased for the more probable group.

In any document retrieval system, the number of documents relevant to a request is always much smaller than the number not-relevant. This establishes two situations:

a. The not-relevant a priori probability is greater than that for relevant which results in decreased misclassification for the not-relevant group.

b. The loss of relevant documents in retrieval increases which reduces the effectiveness of the system.

Therefore, the loss function must be revised to decrease misclassification of relevant documents.

In order to maintain loss as a function of the system, the reciprocal of the a priori ratio is used which moves line AA (fig. 1) to the left and reduces misclassification of relevant documents.

$$\left[(L_R + L_{\bar{R}})/2\right] - \ln(P_{\bar{R}} / P_R)$$

(It is obvious that adding the original loss function produces the same results). This expression defines the location of the decision boundary in terms of its discriminant value.

It is necessary to find this discriminant score in terms of its standardized distance from the relevant and not-relevant group means in order to determine the probabilities of misclassification. This can be done by using the familiar Z score formula which is generally written as

$$Z = \frac{\bar{X} - \mu}{\sigma}$$

Then

$$Z_R = \frac{\left\{\left[(L_R + L_{\bar{R}})/2\right] - \ln(P_{\bar{R}} / P_R)\right\} - L_R}{D}$$

and

$$Z_{\bar{R}} = \frac{\left\{\left[(L_R + L_{\bar{R}})/2\right] - \ln(P_{\bar{R}} / P_R)\right\} - L_{\bar{R}}}{D}$$

The probability of misclassifying a relevant document as not-relevant is given by

$$P_R = p\left[Z_R \geqslant Z\right]$$

and for misclassifying a not-relevant document as relevant

$$P_{\bar{R}} = p\left[Z_{\bar{R}} \geqslant Z\right]$$

both of which can be found by using a normal table of probabilities.

6. RANDOM CLASSIFICATION.

If the documents were to be classified strictly by a random procedure based on a priori probabilities of relevant and not-relevant documents, then the probabilities of correct or incorrect classification would be contingent upon the number of documents in each class. For example, if 10% of the documents are relevant, then the probability of misclassifying a not-relevant document as relevant is .10 and .90 for misclassifying relevant as not-relevant. Each request is described by the number of relevant and not-relevant documents which allows determination of random classification probabilities.

7. RESULTS.

Retrieval effectiveness has been measured by the degree of error introduced by the system as compared to the degree of error which would be introduced by random classification. The system processed two types of request representation.

7.1 REQUEST REPRESENTATION I.

All the documents judged relevant by the request formulators were used to represent the relevance document vector and the remaining documents were used to represent the not-relevant document vector for each request of the 14 which were found to have significant D^2 values (Section 4.3). Results are summarized in Table 1 where

$P(C\bar{R}|R)$ = The probability of misclassifying a relevant document as not-relevant.

$P(CR|\bar{R})$ = The probability of misclassifying a not-relevant document as relevant.

Table 1

PROBABILITIES OF MISCLASSIFICATION FOR 14 REQUESTS USING ALL RELEVANT DOCUMENTS AS THE RELEVANT DOCUMENT VECTOR

	System Results		Random Results	
Requests	$P(C\bar{R}\|R)$	$P(CR\|\bar{R})$	$P(C\bar{R}\|R)$	$P(CR\|\bar{R})$
1. (1)	.004	.151	.954	.046
2. (5)	.002	.087	.964	.036
3. (6)	.004	.063	.912	.088
4. (7)	.001	.046	.985	.015
5. (8)	.001	.007	.933	.067
6. (9)	.002	.055	.974	.026
7. (10)	.001	.033	.970	.030
8. (11)	.001	.029	.979	.021
9. (12)	.001	.001	.995	.005
10. (13)	.001	.013	.990	.010
11. (16)	.001	.023	.964	.036
12. (17)	.002	.072	.979	.021
13. (19)	.001	.001	.985	.015
14. (20)	.001	.001	.979	.021
AVG.	.002	.041	.969	.031

Note: Numbers in parenthesis are original request identifications.

It is obvious the system has performed much better than random classification would allow. On the average, the system misclassifies .2% of the relevant documents as not-relevant and m'sclassifies 4.1% of the non-relevant documents as relevant. Respective average misclassifications due to random procedures are 96.9% and 3.1%.

7.2 REQUEST REPRESENTATION II

Three relevant documents were randomly selected from each of the 19 original requests to represent the relevant document vectors. Eight requests proved significant and these turn out to be a subset of the 14 significant requests used in the previous request representation analysis. Again the system performs much better than chance would allow. On the average, (Table 2) the system misclassifies .07% of the relevant documents as not-relevant and 1.76% of the not-relevant as relevant. Average random errors are 98.5% and 1.5% respectively.

Table 2

PROBABILITIES OF MISCLASSIFICATION FOR 8 REQUESTS USING 3 RANDOMLY SELECTED RELEVANT DOCUMENTS TO REPRESENT THE RELEVANT DOCUMENT VECTOR

	System Results		Random Results	
Requests	$P(C\bar{R}\|R)$	$P(CR\|\bar{R})$	$P(C\bar{R}\|R)$	$P(CR\|\bar{R})$
1. (5)	.0001	.001	.985	.015
2. (7)	.001	.046	.985	.015
3. (10)	.001	.027	.985	.015
4. (11)	.001	.033	.985	.015
5. (13)	.001	.015	.985	.015
6. (17)	.001	.018	.985	.015
7. (19)	.0001	.001	.985	.015
8. (2)	.0001	.001	.985	.015
AVG.	.0007	.0176	.985	.015

8. DOCUMENT RANKS.

In Section 5, it was shown that discriminant scores (L values) could be calculated for every document in the data base within each request by solving the linear discriminant function associated with each document. These scores are projections of each document's geometrical representation on to the discriminant line. Figure 2 shows the range of such scores for a request.

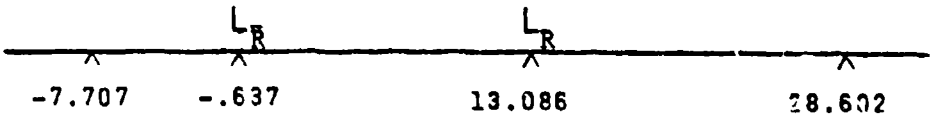

Fig. 2

Range of Discriminant Scores
for a Request

The document with the highest probability of being relevant is represented by the 28.602 discriminant score whereas -7.707 represents the document with the highest probability of being not-relevant. The documents can be sorted in descending order of their discriminant scores and their positions from the top of the list yields their ranks. For all practical purposes, only the known relevant document's ranks need be determined and these are reported in Tables 3 and 4.

Table 3

RELEVANT DOCUMENT RANKS PER REQUEST WHEN ALL RELEVANT DOCUMENTS FORM THE RELEVANCE VECTOR

Requests		Relevant Document Ranks
1.	(1)	1,3,4,5,8,9,10,11,21
2.	(5)	1,2,4,5,6,10,14
3.	(6)	1,2,3,4,5,6,7,8,10,11,12,14,16,22,23,24,31
4.	(7)	1,4,23
5.	(8)	1,2,3,4,5,6,7,8,9,10,11,13,15
6.	(9)	1,2,3,5,13
7.	(10)	1,4,5,6,7,9
8.	(11)	1,2,3,5
9.	(12)	1
10.	(13)	2,3
11.	(16)	1,2,3,4,6,9,14
12.	(17)	1,3,8,12
13.	(19)	1,2,4
14.	(20)	1,2,3,4

Table 4

RELEVANT DOCUMENT RANKS PER REQUEST WHEN 3 RANDOMLY SELECTED DOCUMENTS FORM THE RELEVANCE VECTOR

Requests	Relevant Document Ranks
1. (5)	1,2,4,20,52,68,149
2. (7)	1,4,23
3. (10)	1,4,5,6,7,18
4. (11)	1,2,3,8
5. (13)*	2,4
6. (17)	1,3,10,28
7. (19)	1,2,4
8. (20)	1,2,3,4

*A marginal relevant document was used as the third random document.

8.1 RESULTS OF RANKINGS.

The appearance of many low ranks in Tables 3 and 4 indicates the discriminant function technique does place most of the relevant documents at the upper end of the scale. In Tables 5 and 6 the above data is condensed into intervals of ranks for convenience of analysis.

Table 5

RELEVANT DOCUMENT RANKINGS PER RETRIEVAL INTERVAL WHEN ALL RELEVANT DOCUMENTS FORM THE RELEVANCE VECTOR

	Number of Relevant Documents Retrieved Per Interval					
Requests	1-10	11-20	21-30	31-40	Above 40	Total Relevant
1. (1)	7	1	1			9
2. (5)	6	1				7
3. (6)	9	4	3	1		17
4. (7)	2		1			3
5. (8)	10	3				13
6. (9)	4	1				5
7. (10)	6					6
8. (11)	4					4
9. (12)	1					1
10. (13)	2					2
11. (16)	6	1				7
12. (17)	3	1				4
13. (19)	3					3
14. (20)	4					4
Total	67	12	5	1	0	85
%	78.8	14.1	5.9	1.2	0	

Table 6

RELEVANT DOCUMENT RANKINGS PER RETRIEVAL INTERVAL WHEN 3 RANDOMLY SELECTED DOCUMENTS FORM THE RELEVANCE VECTOR

	Number of Relevant Documents Retrieved Per Interval					
Requests	1-10	11-20	21-30	31-40	Above 40	Total Relevant
1. (5)	3	1			3	7
2. (7)	2		1			3
3. (10)	5	1				6
4. (11)	4					4
5. (13)	2					2
6. (17)	3		1			4
7. (19)	3					3
8. (20)	4					4
Total	26	2	2	0	3	33
%	78.8	6.1	6.1	0	9.1	

8.2 RESULTS OF INTERVAL RETRIEVAL.

High percentages of relevant documents (78.8%) falling into the first interval (ranked among the top 10) demonstrates the discriminant power of the system. Table 5 reveals that 92.9% of the relevant documents are ranked among the top 20 using request representation I. From Table 6 it can be seen that 84.9% are ranked among the top 20.

9. REQUEST FORMULATION.

Although the problem of determining what type of request would be suitable to this system has not been investigated, one form was applied in the course of this experiment.

The original 19 requests were formulated in natural language asking for literature to satisfy a particular need. These requests were transformed into keyword lists of terms selected from the vocabulary on the following basis:

(a) the terms actually appear in the request

(b) subject content suggests particular terms

The geometric representations of the words compiled for each request were mathematically combined to form a request vector. The vector describes a user position in the 39 dimensional space generated by the factor analysis (Section 3.4) and the proximity of this position and document positions reflects similarity of subject content. Euclidean distances between user and documents forms the basis for retrieval and the output is a sorted list of these distances in increasing order with appropriate document identification attached. Table 7 is similar to Tables 5 and 6 in that the document ranks are reported in terms of retrieval intervals.

Table 7

RELEVANT DOCUMENT RANKINGS PER RETRIEVAL INTERVAL USING KEYWORDS AS REQUEST REPRESENTATIONS AND EUCLIDEAN DISTANCE AS RETRIEVAL CRITERION

	Number of Relevant Documents Retrieved Per Interval					
Requests	1-10	11-20	21-30	31-40	Above 40	Total Relevant
1. (1)	4		1	2	2	9
2. (2)	2				2	4
3. (3)	2	1			1	4
4. (4)	3				3	6
5. (5)	1	1			5	7
6. (6)	6	5	2		3	16
7. (7)					3	3
8. (8)	4	1	2	1	5	13
9. (9)		1			4	5
10. (10)	4		1	1		6
11. (11)	3				1	4
12. (12)					1	1
13. (13)	1	1				2
14. (14)	1				4	5
15. (16)	4	1	1	1		7
16. (17)	1		2		1	4
17. (18)	3	3			1	5
18. (19)		1	1		1	3
19. (20)					4	4
Total	39	13	10	5	41	108
%	36.1	12	9.3	4.6	38	

9.1 RESULTS OF KEYWORD RETRIEVAL

Only 36.1% of the relevant documents were ranked among the top 10 whereas 38% were ranked above 40. If only the 14 significant requests are considered (as in Table 5) then 33.3% of the relevant documents receive ranks above 40. These results are not surprising since the requests were not formulated with the multiplicity of keyword occurrences in mind. Opposed to this is the system which has located the documents in the space by giving weight to those words which occur more frequently.

10. DISSEMINATION THRESHOLD SELECTION

In a retrieval system where the output is a ranking of documents according to some relevance criterion, a problem arises in the determination of the number of documents to disseminate to the requestor (user). Sorting and ranking the entire data base can result in wasted computer time; disseminating too many documents to the user can make his selection procedure tedious; and both are impractical if the system were in an on-line mode.

10.1 INTUITIVE THRESHOLD

The use of an intuitive threshold (a constant number of documents based primarily on not giving the user too many documents) which is completely divorced from the relevance criterion could result in low Precision and Recall. Precision is defined as the ratio of relevant documents retrieved to the total documents retrieved (Swets). Recall is the ratio of relevant documents retrieved to the total number of relevant documents in the system (Swets). These measures suggest that the threshold selected have some relationship to the relevance criterion.

10.2 STATISTICAL DECISION BOUNDARY

The most obvious dissemination threshold to use would be the statistical decision boundary described in Section 5.1.

However, it requires estimates of the a priori probabilities of relevant and not-relevant groups of documents which would not be available in practical situations. Consequently, two other dissemination thresholds are calculated and results are compared to the decision boundary (D. B.) results.

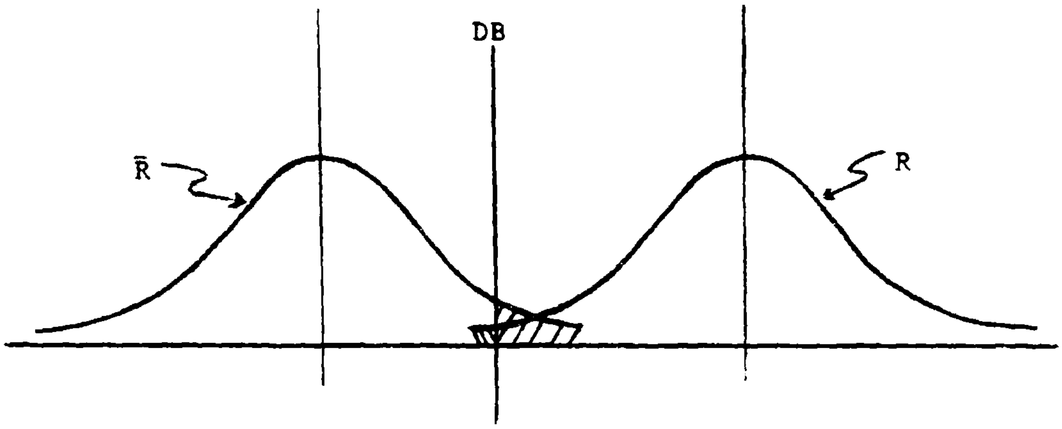

Fig. 3

Statistical Decision Boundary

According to Fig. 3, the use of the statistical decision boundary should result in a small loss of relevant documents (shaded area to the left of the D. B. line) and an introduction of some not-relevant documents (shaded area to the right of the D. B. line). On the average, .2% of the relevant documents would not be included and 4.1% of the not-relevant documents would be included in (Table 1) the group of documents disseminated to the user. The not-relevant set is always much larger than the relevant set which could result in the dissemination of too many not-relevant documents to the user.

It would be desirable to select a cut-off such that the number of not-relevant documents disseminated would be reduced, the loss of relevan' ·· cuments would not be appreciable, and the threshold criterion has some relationship to the relevance criterion. Two such dissemination thresholds (D. T.s) which are distances measured from the mean of the relevant group as a function of the between group variance were calculated as follows:

$$\#1 \text{ D.T.} = L_R - 2\sqrt{D^2}$$

$$\#2 \text{ D.T.} = L_R - 2.5\sqrt{D^2}$$

Tables 8 and 9 show comparisons between retrieval results using D. T.s and the D. B. for request representations I and II respectively.

Table 8.

DOCUMENT DISSEMINATION USING DIFFERENT THRESHOLDS

Request Representation I

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
D.B.	Relevant Retrieved	9	7	16	2	13	5	6	4	1	2	6	4	3	4	82
	Not-Relevant Retrieved	18	20	13	9	2	13	7	9	0	5	7	11	3	0	117
	Relevant Not Retrieved	0	0	1	1	0	0	0	0	0	0	1	0	0	0	3
#1 D.T.	Relevant Retrieved	8	6	13		10	4	5	4	1	2	5	2	2	3	67
	Not-Relevant Retrieved	7	4	7	3	0	5	3	5	0	3	1	4	0	0	42
	Relevant Not Retrieved	1	1	4	1	3	1	1	0	0	0	2	2	1	1	18
#2 D.T.	Relevant Retrieved	9	7	16	2	11	5	6	4	1	2	5	3	2	3	76
	Not-Relevant Retrieved	14	7	10	4	1	9	4	6	0	3	2	6	1	1	68
	Relevant Not Retrieved	0	0	1	1	2	0	0	0	0	0	2	1	1	1	9

Table 9

DOCUMENT DISSEMINATION USING DIFFERENT THRESHOLDS

Request Representation II

		1	2	3	4	5	6	7	8	Total
	Relevant Retrieved	3	2	5	4	2	3	3	4	26
D.B.	Not-Relevant Retrieved	1	9	5	11	7	7	3	0	43
	Relevant Not Retrieved	4	1	1	0	0	1	0	0	7
	Relevant Retrieved	1	2	4	3	2	2	2	4	20
#1 D.T.	Not-Relevant Retrieved	0	3	2	1	5	1	0	0	12
	Relevant Not Retrieved	6	1	2	1	0	2	1	0	13
	Relevant Retrieved	2	2	5	3	2	2	2	4	22
#2 D.T.	Not-Relevant Retrieved	0	4	3	1	5	2	1	0	16
	Relevant Not Retrieved	5	1	1	1	0	2	1	0	11

10.3 RESULTS OF DISSEMINATION THRESHOLD RETRIEVAL

A look at the first request reveals that 27 documents would be disseminated to the user under the decision boundary criterion. Nine of these would be relevant to the request, but 18 would not. Using the #1 threshold, 15 documents are disseminated of which 8 are relevant and 7 are not with one relevant document excluded. Likewise for the #2 threshold, of 23 documents disseminated 9 are relevant and 14 are not with no loss of relevant documents. For this request, the second threshold fulfills the requirements of minimum loss of relevant documents and reduction in the number of not-relevant documents submitted to the user. A few requests do fare better under the decision boundary or the #2 threshold, but on the whole the #1 threshold fulfills the requirements more often.

Table 9 is read in the same manner as Table 8. Overall, the #1 threshold loses 6 relevant documents more than the decision boundary, but reduces the noise level from 43 to 12. In comparison to the #2 threshold, 2 more relevant are lost and 4 less not-relevant documents are disseminated to the user. In light of the fact that the largest number of documents to be disseminated is only 15 (request 11), it would seem that the decision boundary is a more desirable threshold.

10.4 PRECISION AND RECALL

The data in Tables 8 and 9 is converted to Precision and Recall measures (Table 10) to clarify overall analysis of the dissemination thresholds and also to facilitate comparison with other retrieval systems that are evaluated in the same manner.

Table 10

PRECISION AND RECALL MEASURES FOR EACH REQUEST REPRESENTATION

	Request Representation I						Request Representation II					
	Decision Boundary		#1 D.T.		#2 D.T.		Decision Boundary		#1 D.T.		#2 D.T.	
Requests	P	R	P	R	P	R	P	R	P	R	P	R
1. (1)	.33	1.0	.53	.89	.39	1.0						
2. (5)	.26	1.0	.60	.86	.50	1.0	.75	.43	1.0	.14	1.0	.29
3. (6)	.55	.94	.65	.76	.62	.94						
4. (7)	.18	.67	.40	.67	.33	.67	.18	.67	.40	.67	.33	.67
5. (8)	.87	1.0	1.0	.77	.92	.85						
6. (9)	.28	1.0	.44	.80	.36	1.0						
7. (10)	.46	1.0	.63	.83	.60	1.0	.50	.83	.67	.67	.63	.83
8. (11)	.31	1.0	.44	1.0	.40	1.0	.27	1.0	.43	.75	.43	.75
9. (12)	1.0	1.0	1.0	1.0	1.0	1.0						
10. (13)	.29	1.0	.40	1.0	.40	1.0	.22	1.0	.29	1.0	.29	1.0
11. (16)	.46	.86	.83	.71	.71	.71						
12. (17)	.27	1.0	.33	.50	.33	.75	.30	.75	.67	.50	.50	.50
13. (19)	.50	1.0	1.0	.67	.67	.67	.50	1.0	1.0	.67	.67	.67
14. (20)	1.0	1.0	1.0	.75	.75	.75	1.0	1.0	1.0	1.0	1.0	1.0
AVG.	.48	.96	.66	.80	.57	.88	.47	.84	.68	.68	.61	.71

PRECISI $\frac{\text{number of relevant retrieved}}{\text{total retrieved}}$

RECALL = $\frac{\text{number of relevant retrieved}}{\text{total relevant}}$

10.5 RESULTS OF DISSEMINATION IN TERMS OF PRECISION AND RECALL

Overall, the #1 threshold is most desirable for both request representations since it most closely adheres to the conditions of reducing the number of not-relevant documents disseminated to the user without losing too many relevant documents in the process.

For request Representation I, on the average 66% of the retrieved documents are relevant and 80% of the relevant documents are retrieved. The other two thresholds retrieve most of the relevant documents, but of the documents retrieved, about half are not-relevant.

For request Representation II, on the average 68% of the retrieved documents are relevant and 68% of the relevant documents are retrieved. The other two thresholds would retrieve more relevant documents, but would also include more not-relevant documents.

11. SIGNIFICANCE FAILURE INVESTIGATION

In the course of this experiment, an attempt was made to find out why 5 requests failed the statistical significance test. The investigation was centered about the premise that some characteristics of the relevant document sets of these requests differed from the characteristics of the relevant document sets of the requests which were statistically significant.

11.1 SIZE CF RELEVANT SET

Two of the non-significant requests had 4 relevant documents and three had five relevant documents. Three significant requests had 4 relevant documents and one had 5 relevant documents, thus dispelling any suspicion of inadequate requests on the basis of ıelevant set size.

11.2 PROXIMITY OF DOCUMENTS

Since the 39 factors are derived from the occurrences of technical words, documents with similar subject content should be located near each other in the 39 dimensional space. Euclidean distances between each pair of documents in the relevant set can be calculated using respective factor loadings. Average distances can be compared between non-significant and significant requests having the same number of relevant documents. The analysis reveals that there is not any consistent difference between average distances.

11.3 CORRELATIONS

In general correlations are indexes of the degree of agreement between measures on two variables. Factor loadings can be considered as 39 measures and a correlation between two documents can be calculated. These were calculated between each pair of documents in each relevant set of the non-significant and significant requests. These correlations were averaged to allow ease of comparison and again no consistent differences occurred.

The investigation was stopped at this point because it became obvious that a more detailed study would be needed to uncover any disparity between significant and non-significant requests.

12. SUMMARY OF RESULTS AND CONCLUSIONS

The Mahalancbis D^2 - multivariate distance function - is useful in situations where the variables are measured on different scales or where the scales are unknown. Factor analysis has established a subject content space where the axes represent information processing areas. The units of measurement are unknown for such information with makes D^2 appropriate for determining positional relationships between groups of documents. The fact that these relationships can be statistically tested lends to the usefulness of such a measure for evaluating retrieval effectiveness. As for the linear discriminant function, it provides numerical values for document positions which in turn allows ranking them for purposes of determining the most probable relevant documents to the most probable not-relevant documents. The linear function maximizes separation between relevant and not-relevant groups which tends to improve the accuracy of a retrieval system that uses information of this nature. The feature that makes Mahalanobis D^2 and the linear discriminant function so important is that together they allow the measurement of system error without going through the mechanics of retrieval.

Two types of request representations were used to obtain results for evaluating retrieval effectiveness in terms of the degree of error introduced by the system as opposed to random classification.

Using request Representation I, the system operates with an average error of .2% when classifying relevant documents and 4.1% when classifying not-relevant documents. Random classification on the average produces errors of 96.9% and 3.1% respectively. Using request Representation II, average error rates are .07% and 1.76% respectively and for random classification they are 98.5% and 1.5% respectively. These results imply the 39 factors derived from the statistical associations between technical words adequately describe the classification space so that effective retrieval can be accomplished.

The discriminant power of the system was tested by ranking discriminant scores associated with each document from largest to smallest and calculating percentages of relevant documents occurring in retrieval intervals. This was done for each request under the two types of request representations. For request Representation I, 78.8% of the relevant documents are ranked among the top 10 and 92.2% among the top 20. Under request Representation II, the results are 78.8% and 84.9% respectively.

One method of formulating requests for retrieval was tried and resulted in only 48.1% of the relevant documents having ranks among the top 20, proving keyword type of requests per se are not suitable to this system. The type of request that is suitable is not evident at this time and warrants further study.

A dissemination threshold was sought which would reduce the number of not-relevant documents included in retrieval without dropping an appreciable number of relevant documents. It was found that the #1 D. T. complies with this restriction. The data was transformed by Precision and Recall measures to facilitate comparisons with other retireval systems. For request Representation I, #1 D. T. results in 66% of the retrieved documents being relevant and 80% of the relevant documents being retrieved. For request Representation II, both are 68%. The system has achieved higher Precision and Recall than most retrieval systems have that are in use today.

Three types of tests were performed on the 5 requests which failed the statistical significance test to determine if they possessed characteristics different from the 14 significant requests. The findings were negative and suggest that a more detailed study is necessary to either validate these results or expose any underlying differences.

Overall, results have shown that an effective automatic retrieval system using statistical word associations can be built which would perform the retrieval task at a higher level than most systems do today. Multivariate statistics has application to systems of this type, both in improving discrimination between relevant and not-relevant documents and as a measure of retrieval effectiveness.

BIBLIOGRAPHY

Anderson, T.W., Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc., New York, 1958.

Cooley, W.W. and Lohnes, P.R., Multivariate Procedures for the Behavorial Sciences, John Wiley & Sons, Inc., New York, 1962.

Harmon, H.H., Modern Factor Analysis, University of Chicago Press, Chicago, 1960.

Hays, W.L., Statistics for Psychologists, Holt, Rinehart and Winston, New York, 1963.

Lawley, D.N. and Maxwell, A.E., Factor Analysis as a Statistical Method, Butterworths, London, 1963.

Li, C.C., Introduction to Experimental Statistics, McGraw-Hill Book Company, New York, 1964.

Ossorio, P.G., "Classification Space Analysis," Technical Report RADC-TR-64-287, October 1964, University of Colorado, Contract AF 30(602)-3342, AD 608 034.

Salton, G., Information Storage and Retrieval, Scientific Report No. ISR-13 to National Science Foundation, December 1967.

Swets, J.A., "Information Retrieval Systems," Science Vol. 141, July 1963.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Rome Air Development Center (EMIDB) Griffiss Air Force Base, New York 13440	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP N/A

3. REPORT TITLE

STATISTICAL INFORMATION RETRIEVAL SYSTEM

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
In-House Report

5. AUTHOR(S) (First name, middle initial, last name)

Nicholas M. DiFondi

6. REPORT DATE October 1969	7a. TOTAL NO. OF PAGES 45	7b. NO. OF REFS 9
8a. CONTRACT OR GRANT NO. N/A b. PROJECT NO. 4594 c. Task 459401 d.	9a. ORIGINATOR'S REPORT NUMBER(S) RADC-TR-69-382 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Rome Air Development Center (EMIDB) Griffiss Air Force Base, New York 13440

13. ABSTRACT

An information retrieval system was developed using technical word occurrences as a basis for classification.

A set of words, designated a vocabulary, was selected from the middle range of frequency listing of words occurring in an experimental sample of 94 documents. The selection produced 115 non-function words with technical definition that did not allow ambiguous usage and they were assigned one of eighty concept numbers. The frequencies of these concepts served as data for factor analysis and 39 factors were extracted to represent the orthogonal axes of a geometric subject-content space. The locations of concepts in this space were used to locate the geometric position of documents according to their frequencies in the documents.

The total of 194 documents was used in the measuring of system effectiveness. The Mahalanobis D^2 function provided a statistical measure of the separation between relevant and not-relevant groups in the space. Linear discriminant functions were solved to maximize between group differences and Fisher's variance ratio was used to test the significance of group separation. Empirical and theoretical probabilities of misclassification were compared and system error on the average was .2% for relevant and 4.1% for not-relevant documents. Theoretical errors were 96.9% and 3.1% respectively. The small system errors validated the accuracy of the 39 dimensional subject-content space.

(continued)

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information Retrieval Classification Subject Indexing						

Abstract (continued)

Requests formulated for a previous experiment using the same data base were processed. Precision and Recall measures were calculated and on the average 66% Precision and 80% Recall were attained with one of three dissemination thresholds.

Overall analysis of the results supports the theory that statistical data about word occurrences is sufficient to accurately represent documents relative to their subject content.

UNCLASSIFIED
Security Classification

END